## **Predicting the “Re-admission possibility of a patient into the hospital”**

**Purpose:**

A leading hospital in the US is suddenly seeing increase in the patient readmission in less than 30 days. This is serious concern for the hospital as it may indicate insufficient treatment or diagnosis when the patient was admitted first and later released under clean bill of health.

Not only the image of hospital as healthcare provider is compromised, this is also increased cost to the entire medicare ecosystem in form of increased insurance claims.

**Problem Statement :**

it is in Hospitals interest to support their diagnosis by a better predictive model which you are going to build.

**Objective**:

To classify the patients treated by this hospital into two primary categories:

* Readmitted within 30 days
* Not readmitted

**Data Set and Attribute Information:**

Following are the datasets provided and the brief description about the datasets.

**Attribute Information:** The dataset depicts the details of the patients, like, treatments that each patient has underwent, patient previous medical information, procedures (lab, procedures, diagnosis, medications)underwent by ,and their demographic info etc**.**

**Information about attributes in each data set:**

**Train Data :**

* Race: For this attribute, missing values are denoted as “?".
* Weight: For this attribute, missing values are denoted as “?”.

|  |  |  |
| --- | --- | --- |
| Attribute | No. of levels | Type |
| patientID | 34650 | Categorical |
| race | 5 | Categorical |
| age | 10 | Range |
| weight | 8 | Range |
| readmitted | 2 | Categorical |
| gender | 2(Male, Female) | Categorical |

**Train\_HospitalizationData:**

* payer\_code: For this attribute, missing values are denoted as “?”
* medical\_specialty: For this attribute, missing values are denoted as “?”.

|  |  |  |
| --- | --- | --- |
| Attribute | No. of levels | Type |
| AdmissionID | 34650 | Categorical |
| patientID | 34650 | Categorical |
| Admission\_date | Dates(2014,15,16) | Date |
| Discharge\_date | Dates(2014,15,16) | Date |
| admission\_type\_id | 8 | Numeric |
| discharge\_disposition\_id | 28 | Numeric |
| admission\_source\_id | 25 | Numeric |
| payer\_code | 17 | Categorical |
| medical\_specialty | 66 | Categorical |

**Train\_Diagnosis\_TreatmentData :**

|  |  |  |
| --- | --- | --- |
| Attribute | No. of levels | Type |
| patientID | 34650 | Categorical |
| num\_lab\_procedures | 132 | Numerical |
| num\_procedures | 6 | Numerical |
| num\_medications | 81 | Numerical |
| num\_diagnoses | 16 | Numerical |
| diagnosis\_1 | 664 | Numerical |
| diagnosis\_2 | 681 | Numerical |
| diagnosis\_3 | 726 | Numerical |
| max\_glu\_serum | 4 | Categorical |
| A1Cresult | 4 | Categorical |
| metformin | 4 | Categorical |
| repaglinide | 4 | Categorical |
| nateglinide | 4 | Categorical |
| chlorpropamide | 4 |  |
| glimepiride | 4 | Categorical |
| acetohexamide | 1 | Categorical |
| glipizide | 4 | Categorical |
| glyburide | 4 | Categorical |
| tolbutamide | 2 | Categorical |
| pioglitazone | 4 | Categorical |
| rosiglitazone | 4 | Categorical |
| acarbose | 3 | Categorical |
| miglitol | 2 | Categorical |
| troglitazone | 2 | Categorical |
| tolazamide | 2 | Categorical |
| insulin | 4 | Categorical |
| glyburide.metformin | 4 | Categorical |
| glipizide.metformin | 2 | Categorical |
| metformin.rosiglitazone | 2 | Categorical |
| metformin.pioglitazone | 2 | Categorical |
| change | 2 | Categorical |
| diabetesMed | 2 | Categorical |

**Data Analysis and Pre processing:**

**Merged train Data :** This dataset has 34650 rows and 42 columns in train set and

**Merged test Data** 14630 rows and 42 columns in test set.

**From Train data:**

* The first column in Train data " **patient**ID " represents unique patient ID. This column is present in the other two data sets as well and was used for merging the datasets.
* Race column has 6 levels ( Caucasian, AfricanAmerican, Other, Hispanic, Asian)
* Age has ranges of levels. Ex. [10-20]
* Weight describes the weight bracket a patient falls into.

**Missing values**:

1079 missing values and 432 missing values for the feature race, 33592 missing values and 14149 missing values for the feature weight in the train and test dataset respectively represented as ?.

* Imputed NAs using central imputation.
* Weight column was removed from the final merged data set.

**From Hospitalization data:**

**Missing values** :

Train Demographics data has 14719 missing values and 6283 missing values for the feature payer\_code, 16394 missing values and 7000 missing values for the feature medical\_speciality in the train and test dataset respectively represented as ?.

* Imputed NAs using central imputation.
* Dropped payer\_code and medical\_speciality

**From Diagnosis data:**

**Missing values:**

Train: Diagnosis\_1 = 6, diagnosis\_2= 179, diagnosis\_3 = 681

Test: Diagnosis\_1 = 3, diagnosis\_2= 65, diagnosis\_3 = 286

* Initially dropped diagnosis\_1, diagnosis\_2, diagnosis\_3, chlorpropamide, acetohexamide, tolbutamide, miglitol, troglitazone, tolazamide, glipizide.metformin, metformin.rosiglitazone, metformin.pioglitazone because of level imbalance and large number of levels.
* **Feature Generation :**
* In one approach from the admission\_date and discharge\_date , Tenure was created.
* Tried binning diagnosis\_1,diagnosis\_2 and diagnosis\_3 in one approach.

**Under-sampling:**

**Oversampling** and **undersampling** in data analysis are techniques used to adjust the class distribution of a [data set](https://en.wikipedia.org/wiki/Data_set) (i.e. the ratio between the different classes/categories represented).

Oversampling and undersampling are opposite and roughly equivalent techniques. They both involve using a [bias](https://en.wikipedia.org/wiki/Bias) to select more samples from one class than from another.

The usual reason for oversampling is to correct for a bias in the original dataset. One scenario where it is useful is when training a classifier using labelled training data from a biased source, since labelled training data is valuable but often comes from un-representative sources.

For example, suppose we have a sample of 1000 people of which 66.7% are male. We know the general population is 50% female, and we may wish to adjust our dataset to represent this. Simple *oversampling* will select each female example twice, and this copying will produce a balanced dataset of 1333 samples with 50% female. Simple *undersampling* will drop some of the male samples at random to give a balanced dataset of 667 samples, again with 50% female.

**Correlation coefficient:**

A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variables with a known distribution.

Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics. They all assume values in the range from −1 to +1, where +1 indicates the strongest possible agreement and −1 the strongest possible disagreement. As tools of analysis, correlation coefficients present certain problems, including the propensity of some types to be distorted by outliers and the possibility of incorrectly being used to infer a causal relationship between the variables.

**Correlation between variables**:

race, gender, age, weight, Admission\_date, Discharge\_date,

admission\_type\_id, discharge\_disposition\_id, admission\_source\_id,

payer\_code, medical\_specialty, num\_lab\_procedures,

num\_procedures, num\_medications, num\_diagnoses, diagnosis\_1,

diagnosis\_2, diagnosis\_3, max\_glu\_serum, A1Cresult, metformin,

repaglinide, nateglinide, chlorpropamide, glimepiride,

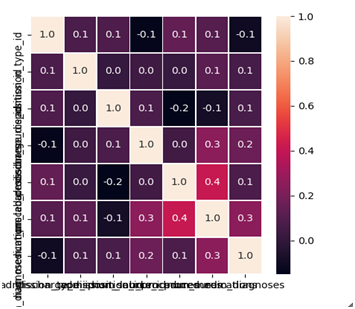
acetohexamide, glipizide, glyburide, tolbutamide,

pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone,

tolazamide, insulin, glyburide.metformin, glipizide.metformin,

metformin.rosiglitazone, metformin.pioglitazone, change,

diabetesMed

****

**Splitting the data into train and validation datasets:**

* Train set: 90%
* Validation set:10%

**Error Metrics: Recall and accuracy**

Recall is important because its important to rightly predict the readmission rate .(True positive rate should be more out of actual predictions.)

**5. Model Building:**

Following different classifying techniques were used. Short information about each of the classifying techniques namely

* Multinomial logistic regression
* Naive Bayes
* Decision tree classifier
* Random forest
* SVM-Classifier
* Stacking

***Multinomial logistic regression*** is used when the dependent variable  in question is nominal and for which there are exactly two categories.

precision recall f1-score support

NO 0.62 0.64 0.63 479

Within30days 0.62 0.60 0.61 473

avg / total 0.62 0.62 0.62 952

***Naïve Bayes***is a classificationtechnique based on Bayes Theorem with an assumption of independence among predictors. In simple terms, a Naive Bayes classifier assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature .

**Accuracy** 0.1743147173680739

**Recall** 0.97

|  |  |
| --- | --- |
| **Dataset** | **Accuracy** |
| validation | 0.1743147173680739 |

***Random forests*** improve predictive accuracy by generating a large number of bootstrapped trees (based on random samples of variables), classifying a case using each tree in this new "forest", and deciding a final predicted outcome by combining the results across all of the trees (an average in regression, a majority vote in classification).

**Hyperparameters:**

{'max\_depth': 5, 'min\_samples\_split': 30, 'n\_estimators': 40}

precision recall f1-score support

NO 0.62 0.62 0.62 479

Within30days 0.61 0.61 0.61 473

avg / total 0.61 0.61 0.61 952

***Classification Trees*** for short is a term introduced to refer to Decision Tree algorithms that can be used for classification or regression predictive modeling problems. Classically, this algorithm is referred to as “decision trees”, but on some platforms like python they are referred to by the more modern term DecisiontreeClassifier. The DecisiontreeClassifier algorithm provides a foundation for important algorithms like bagged decision trees, random forest and boosted decision trees.

**Hyperparameters:**

(criterion = "entropy",max\_depth=8,min\_samples\_leaf = 1,min\_samples\_split = 130)

precision recall f1-score support

NO 0.64 0.45 0.52 479

Within30days 0.57 0.74 0.64 473

avg / total 0.60 0.59 0.58 952

***SVM -classifier*** Support Vector Machine is a supervised **machine learning** algorithm which can be used for both classification or regression challenges. However, it is mostly used in classification problems. It is a frontier which best segregates the two classes (hyper-plane/ line)

* rbf Svm
* Search: Random
* Number of cross validation splits: 5

precision recall f1-score support

NO 0.58 0.67 0.62 479

Within30days 0.60 0.51 0.55 473

avg / total 0.59 0.59 0.58 952

***Stacked model***

Ensemble methods are commonly used to boost predictive accuracy by combining the predictions of multiple machine learning models. The traditional wisdom has been to combine so-called “weak” learners. However, a more modern approach is to create an ensemble of a well-chosen collection of strong yet diverse models.Building powerful ensemble models has many parallels with building successful human teams in business, science, politics, and sports. Each team member makes a significant contribution and individual weaknesses and biases are offset by the strengths of other members.The simplest kind of ensemble is the unweighted average of the predictions of the models that form a model library.

precision recall f1-score support

0 0.55 0.52 0.54 479

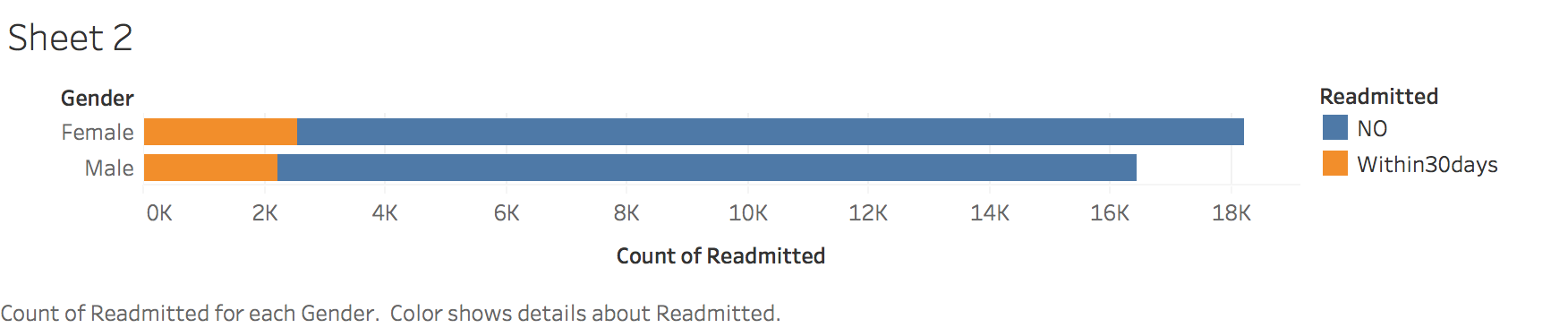
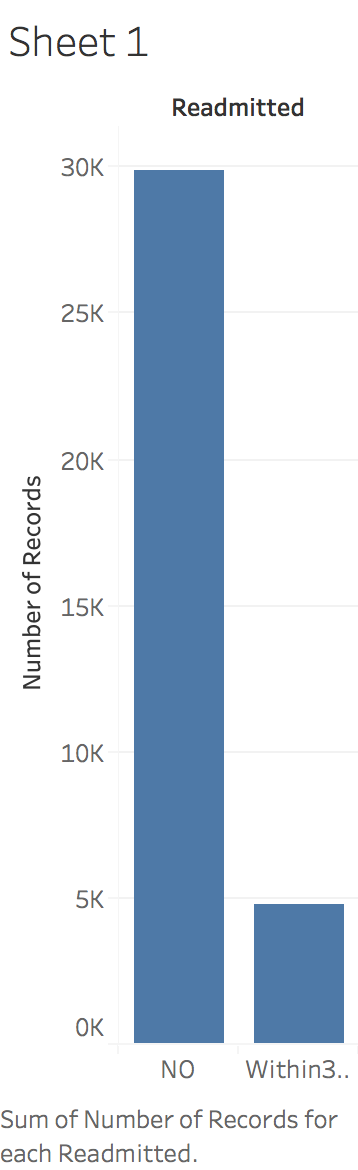
1 0.54 0.57 0.56 473

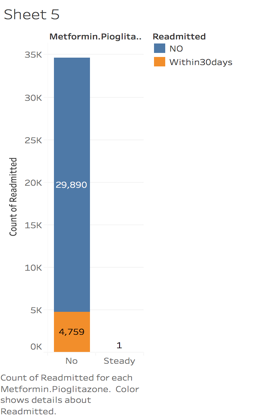
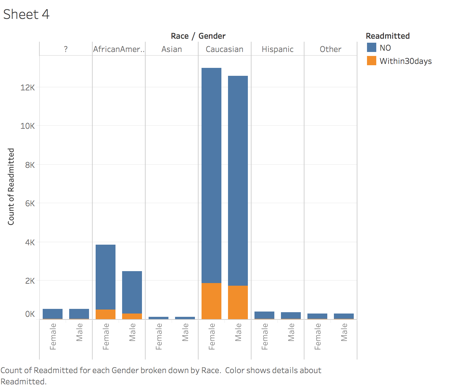
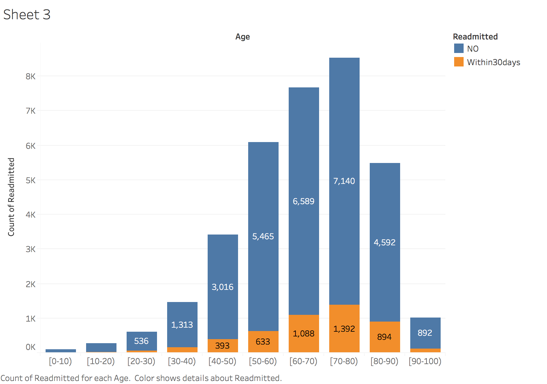
avg / total 0.55 0.55 0.55 952

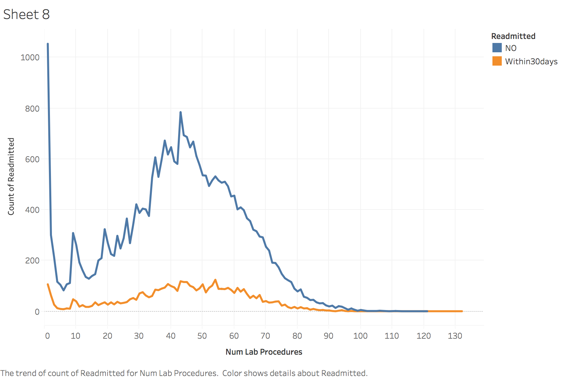
**Model Selection:**

**Best model:**

* DecisionTreeClassifier is giving the best recall and accuracy

****

****

****

**Conclusion:**

This assessment has given me a fair understanding of how to deal with class imbalanced in data, feature creation, feature selection, tweaking with parameters, hypermeter selection, grid search functioning and model stacking. The whole task proved that simple model is always better over complex one. Inferences made from this simple model should help the hospital reduce the number of readmitted patients within 30 days.